一般論文

Peer-Reviewed

BioVL2 データセット：生化学分野における一人称視点の実験映像への言語アノテーション

西村　太一[†]・迫田航次郎[†]・牛久　敦[†]・橋本　敦史[††]・奥田奈津子[†††]・小野富三人[†††]・亀甲　博貴[††††]・森　信介[††††]

本論文では，生化学分野における一人称の実験映像データセットである BioVL2 データセットを提案する．BioVL2 データセットは生化学における 4 種類の基本的実験に対し，それぞれ 8 動画撮影した合計 32，総時間 2.5 時間の映像からなるデータセットである．各映像はプロトコルと紐づいており，言語アノテーションとして (1) 視覚と言語の対応関係のアノテーション，(2) プロトコル中に現れる物体の矩形アノテーションの 2 種類のアノテーションを付与している．構築したデータセットの応用例として，本研究では実験映像からプロトコルを自動生成する課題に取り組んだ．定量的，定性的な評価の結果，開発した手法はフレームに映っている物体名をそのままプロトコルとして出力する弱いベースラインと比較して，適切なプロトコルを生成できることを確認した．なお，BioVL2 データセットは研究用途に限定してデータセットを公開する予定である *.

キーワード：生化学分野，プロトコル，視覚と言語の融合研究

BioVL2: An Egocentric Biochemical Video-and-Language Dataset

Taichi Nishimura[†], Kojiro Sakoda[†], Atsushi Ushiku[†], Atsushi Hashimoto[††], Natsuko Okuda[†††], Fumihito Ono[†††], Hirotaka Kameko[††††] and Shinsuke Mori[††††]

In this study, we propose an egocentric biochemical video-and-language dataset called BioVL2 comprising eight videos for each of four experiments, with a total duration of 2.5 hours for all 32 samples. Each video corresponds to a protocol and two types of linguistic annotations are provided: (1) video-and-text alignment and (2) bounding boxes linked to objects in the protocol. As an application of the BioVL2 dataset, we consider the task of generating a protocol from an experimental video. Our experimental results show that the proposed system can generate better protocols than a

[†] 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University
[††] オムロン サイニックエックス株式会社, OMRON SINIC X Corporation
[†††] 大阪医科薬科大学医学部生命科学講座生理学教室, Department of Physiology, Division of Life Sciences, Faculty of Medicine, Osaka Medical College
[††††] 京都大学学術情報メディアセンター, Academic Center for Computing and Media Studies, Kyoto University

* 本論文は (Nishimura et al. 2021b) で発表した BioVL データセットの拡張版である．拡張内容については，1 節の最終段落を参照されたい．

weak baseline designed to output objects appearing in the video frames. The BioVL2 dataset will be released for research purposes only.

Key Words: *Biochemical Domain, Protocols, Vision-and-Language*

1　はじめに

科学は再現性の危機に瀕している．生化学や生命科学などの薬品を用いた化学実験を行う研究分野においては，75%から 80%以上の研究者が他の研究者の実験結果を再現することができなかった経験があると報告している (Baker 2016)．化学実験で再現性を担保する上で鍵となるのがプロトコルである．プロトコルは人がある実験を再現するために必要な操作を時系列順に記述した文書である（図 1）．プロトコルには，試薬や装置などの操作対象の物体名と，対応する操作方法が動詞で，実験を再現するのに必要十分な記述がされている[1]．加えて，必要であれば物体の量や，操作する時間，あるいは操作の様態が副詞で記述されていることもある．例え

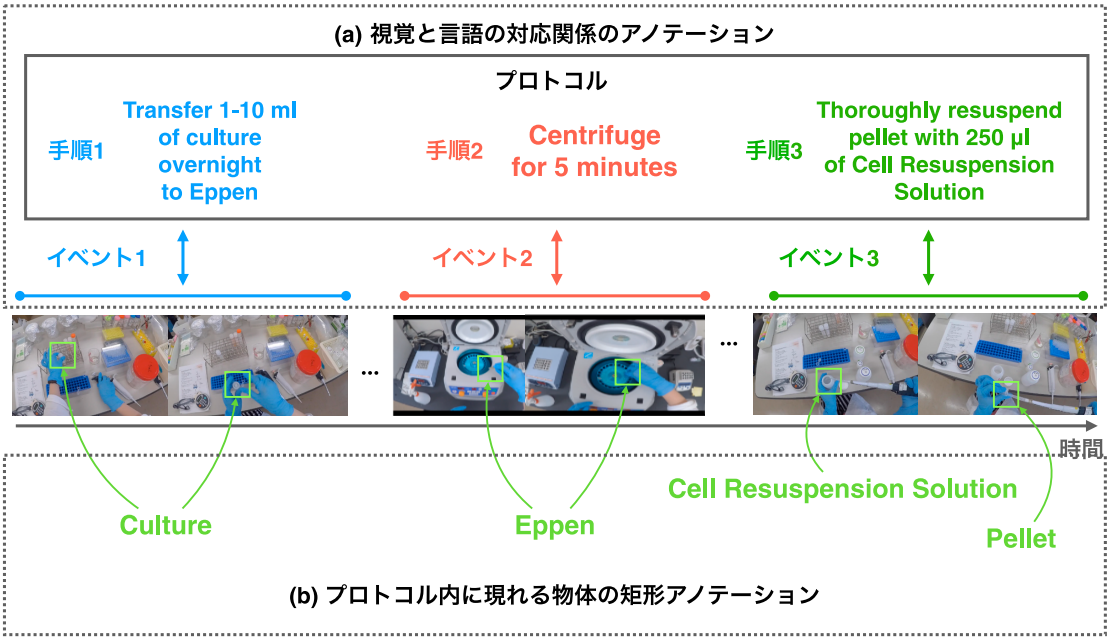

図 1　BioVL2 データセットの概要．(a) の視覚と言語の対応関係のアノテーションに加え，(b) プロトコル中に現れる物体の矩形アノテーションを付与している．

[1] 自明である物体名に関しては省略されることもある．例えば，図 1 の手順 2 では手順 1 の成果物を指しているが，明示的に記述してはいない．

ば，図 1 の手順 3 の “Thoroughly resuspend pellet with 250 μL of Cell Resuspension Solution” では，pellet，Cell Resuspension Solution という物体名の記述があり，resuspend という操作方法が動詞で記述されている．加えて，Thoroughly という副詞や 250 μL という量に関する記述もある．こうしたプロトコルに従って実行することで，理想的には実験を再現することができるはずだが，操作に抜け漏れがあったり，操作の詳細が記述されていなかったりといった問題があると，他の研究者が実験を再現することが困難になる．

こうした再現性の危機に関する問題に対する有望な解決となりうるのが，視覚と言語の融合研究である．例えば，撮影した実験映像とプロトコルの組から，映像の操作シーンとプロトコルの各手順の対応関係を推定できれば，手順ごとに視覚的に操作を確認できる．あるいは，作業映像を入力としてプロトコルを自動生成できれば研究者がプロトコルを書く負担を軽減することができる．このように，化学実験を対象とした視覚と言語の融合研究は実験プロトコルの参照時と作成時の両方の負担を軽減し，実験再現性の向上に資するであろう．

こうした有用性はあるものの，実験映像を対象とした視覚と言語の融合研究の数は多くない (Naim et al. 2014, 2015)．その原因の 1 つに，実験映像を撮影し公開することが困難な点にある．現に，Naim らの研究で利用しているデータセットは公開されていない．そのため，我々はこの目標に向けた第一歩として，生化学分野を対象として実験映像を収集し，言語アノテーションを付与した BioVL2 データセットを構築し研究コミュニティに公開する（図 1）．具体的には以下の 2 種類のアノテーションを作業映像に付与する．

(1) **視覚と言語の対応関係のアノテーション．**プロトコルを動詞ごとに分割した文のそれぞれに対して（本論文ではこれを特に**手順**と呼ぶ），映像の中で手順が実施されている区間（以下，**イベント**と呼ぶ）を付与する．このアノテーションは従来の視覚と言語の融合研究 (Zhou et al. 2018a; Krishna et al. 2017) と同様であり，映像キャプショニング (Xu et al. 2016; Nishimura et al. 2021a) や映像と視覚の対応関係の推定 (Naim et al. 2014, 2015) などの応用研究に活用できる．

(2) **プロトコル内に現れる物体の矩形アノテーション．**映像中の各フレームごとに，プロトコル中の物体が写っていて，かつ実験者の手と接触があった場合に物体の矩形情報を付与する．これにより，映像中の空間的な分析（例：何が写っているか，どういう状態か）や実験者の動作分析が可能になる．また，前述のアノテーションと合わせてプロトコル中の物体名と映像中の物体との対応関係の推定 (Zhou et al. 2019b) などの応用研究にも利用できる．

これらのアノテーションの付与を行うことで，映像からのプロトコル生成や手順を入力としたシーン検索が可能となる．こうした検索が行えると，初学者に対する教育効果や作業補助が期待でき，実験の再現性の向上につながる．また，データがさらに集まるようになれば，最終的にはプロトコルからのロボット操作などのより挑戦的，かつ有用性が高い課題にも取り組む

ことが可能になる．本研究で提案する BioVL2 データセットはこうした生化学実験を対象とした言語と視覚の融合研究への第一歩である．

BioVL2 データセットの収集において意識した設計は，一人称視点のカメラを用いることで，研究者への撮影の負担を最小限にしたことである．実験の度に大掛かりな撮影環境を構築していては，日々実験を行う研究者らは撮影に負担を感じ，結果データセットのサイズはスケールしない．研究者らが自ら撮影に取り組めるように，できるだけ研究者への負担が少ない設計を考える必要がある．この点で，三人称カメラは撮影の度に広範な実験空間をカバーするのに複数台の設置が必要で，故障のリスクが高くなる他，同時撮影などの手間が発生する．一人称カメラは広範な実験空間をカバーしつつも，生化学分野の研究者が手軽に撮影可能である．これが一人称カメラを用いた理由である．こうして撮影を行った結果，全 32 の実験映像とそのアノテーションからなるデータセットを構築した．

得られた BioVL2 データセットを用いて，その応用として本論文では実験映像からプロトコルを生成する課題に取り組む．実験映像の数は他の映像キャプショニングのデータセット (Krishna et al. 2017; Zhou et al. 2018a; Xu et al. 2016) と比較すると少なく，こうした課題で提案されている End-to-end な深層学習モデルを本課題に直接適用することは困難である．そのため，本研究では，Ushiku ら (Ushiku et al. 2017) によって提案された手順書生成モデルを活用する．このモデルは本研究と同様，少量の料理映像（20 映像）に対して適用できるように外部リソースを活用しながら学習できるよう設計されている．このモデルにいくつかの改良を施し，BioVL2 データセットの実験映像からプロトコル生成を生成する課題に取り組む．定量的，定性的評価の結果，モデルは弱いベースラインと比較して，適切なプロトコルを生成できることを確認する．

本論文で述べる BioVL2 データセットは (Nishimura et al. 2021b) にて発表した BioVL データセットの拡張である．具体的には，(1) 映像の数を 16 から倍の 32 へ増加させたこと，(2) 映像への矩形アノテーションを追加で行ったことの 2 点の拡張を行った．さらに，(Nishimura et al. 2021b) では行わなかった，実験映像からプロトコルを生成する課題に取り組んだことも本研究の追加の貢献である．BioVL データセットと同様，BioVL2 データセットは研究用途に限り公開する予定である[2]．

2 関連研究

関連研究を大きく分けて 2 つの観点から説明する．第一に，作業映像と言語アノテーションの観点から，第二に，生物学の諸分野を対象とした自然言語処理，および視覚と言語の融合研究の観点から本研究の位置付けについて述べる．

[2] https://github.com/misogil0116/BioVL2

2.1 作業映像と言語アノテーションからなるデータセット

2.1.1 Web 上の作業映像を対象とした研究

ある目的に向かい処理を進める行動系列を理解するために，作業映像を対象とした研究は近年活発に取り組まれている．作業映像の理解の手がかりとして，言語情報をアノテーションする方法がよく採用され，様々なデータセットが提案されてきた．数ある分野（例：料理，家具の組み立て）の中で，最も注目を集めてきたのは料理分野である．この分野は Web 上でデータを集めやすく，物体や動作の種類が多様であることから，長年にわたり研究者の注目を集め続けてきた (Nishimura et al. 2021a; Zhou et al. 2019b; Wang et al. 2021)．料理分野で提案された映像と言語アノテーションからなるデータセットのうち，最も大規模なデータセットは YouCook2 (Zhou et al. 2018a) データセットである．このデータセットは YouTube 上の料理映像を 89 のカテゴリに分けて収集し，映像に対し料理を達成する上で重要なイベントのアノテーションを行い，各イベントに対し手順を付与したデータセットとなっている．そして，映像の密キャプショニング (Krishna et al. 2017; Zhou et al. 2018b)，作業映像における質問応答 (Wang et al. 2021)，作業映像からの手順書生成 (Nishimura et al. 2021a) など様々な研究におけるベンチマークデータセットとしてもよく用いられている．また，料理分野に続き，メイクアップ映像を YouTube から収集し同様のアノテーションを付与したデータセットである YouMakeup (Wang et al. 2019) も提案されている．

特定の分野に限定せず，作業映像一般を収集し言語アノテーションを付与する取り組みも行われている．Howto100M (Miech et al. 2019) はその中でも最も大規模なデータセットであり，1 億の作業映像と言語の組からなるデータセットである．なお，Howto100M に付与している言語アノテーションは，映像に付与しているナレーションを言語アノテーションとして自動的に収集し付与したものである．

2.1.2 一人称視点の作業映像を対象とした研究

Web 上の映像の多くは視聴者のために編集されていることが多く，作業を行う上で必要な行動の全てが映像中に現れるわけではない．例えば，冷蔵庫から材料を取り出すシーンや，ミキサーで混ぜるなどの時間がかかるシーンは編集で切り取られやすい．こうしたシーンも含めた作業行動全体の理解のために，近年，一人称視点の未編集映像を収集する取り組みが行われてきた．料理分野における代表的なデータセットとして，EPIC-KITCHEN データセット (Damen et al. 2018) がある．このデータセットは，32 の参加者の料理を行う過程を一人称視点のカメラを使って撮影したもので，合計 100 時間，700 映像からなり，各映像には全ての人間の動作が言語情報でアノテーションされている（例：冷蔵庫を開ける，じゃがいもを切るなど）．他にも，おもちゃの車の組み立てを行うためのデータセットとして Assembly101 データセット (Sener et al. 2022) が提案されている．このデータセットは合計 513 時間，4,321 映像からなるデータ

セットであり，EPIC-KITCHENS データセットと同様に動作情報のアノテーションとして言語アノテーションが行われている．

BioVL2 データセットは，生化学分野における一人称視点の作業映像と言語アノテーションからなるデータセットであると位置付けられる．前述したデータセットと比較すると撮影した映像数は少ないが，今後の生化学分野における実験映像理解の研究において有用である．

2.2 生物学の諸分野を対象とした自然言語処理，および視覚との融合研究

生物学およびその諸分野（生化学，分子生物学）において，再現性の観点からも人工知能技術を適用することのニーズは高い（例：プロトコルから実験映像を検索，プロトコルからロボットの実験実行）．以下では，自然言語処理，および視覚との融合研究に関する先行研究を挙げ，本研究の位置づけを明確にする．

2.2.1 自然言語処理

生命科学を対象とした代表的なデータセットとして，GENIA コーパス (Kim et al. 2003) がある．これは，生命科学論文のアブストラクトに対して，品詞情報や構文木，固有表現などの計 6 種類のアノテーションが付与されたデータセットである．Kulkarni ら (Kulkarni et al. 2018) は，生物学のプロトコルを処理するためのデータセットである Wet Lab Protocol データセット（WLP データセット）を提案した．生物学のプロトコルを機械可読な動作グラフ表現 (Kiddon et al. 2015) に変換することがこのデータセットの目的であり，`https://www.protocols.io/` から収集した全 622 のプロトコルについて，固有表現のアノテーション（bio-Named Entity，以下 b-NE と呼ぶ），加えてそれらの関係性（例えば，ある薬品をどのように操作するといった述語項構造のヲ格に当たる関係）を閉路のない有向グラフで表現するアノテーションを付与している．そして，このアノテーションの根のノードがプロトコルの最終成果物を表している．EMNLP2020 併設のワークショップ The 6th Workshop on Noisy User-generated Text (W-NUT) では，このデータセットをもとに共有タスクとして取り組まれ，固有表現認識および関係抽出の課題において多くの手法が提案されている (Knafou et al. 2020; Singh and Wadhawan 2020; Sohrab et al. 2020)．本研究において，WLP データセットの b-NE タグのアノテーション基準は BioVL2 データセットのプロトコル中の動詞のタグ付けや，物体名の抽出などにおいて参考にしている．また，節 4 において，手順書生成のための文生成モデルの事前学習においても WLP データセットを活用している．

2.2.2 言語と視覚の融合研究

自然言語処理技術のみでの発展と比較すると，言語と視覚の融合研究の数は多くはない．Naim ら (Naim et al. 2014, 2015) はプロトコルと実験映像を入力としてプロトコル中の手順と動画の

区間について対応関係を獲得する課題に対し，(Naim et al. 2014) では IBM モデル (Brown et al. 1993) をもとに，(Naim et al. 2015) では条件付き確率場 (Lafferty et al. 2001) をもとにして教師なしで学習する手法を提案している．Naim らの研究との本研究の差分は主に以下の 3 点にある．第一に，前述した研究ではデータセットを公開していないが，BioVL2 データセットは公開する点．第二に，プロトコルの種類数，映像数ともに Naim らのそれを上回る点（BioVL2 データセットが 4 種類 8 動画に対し，Naim らのデータセットは 3 種類 2 動画）．第三に，本研究ではデータセットの応用課題として，言語と視覚の対応関係の獲得ではなく，実験映像からの手順書生成を行う点である．

3　BioVL2 データセット

本節では，構築した BioVL2 データセットについて説明する．まず，データセットの構築方法について述べ，次に統計情報およびアノテーションの一致率を報告する．

3.1　データセットの構築

3.1.1　実験映像の撮影

撮影環境について． 1 名の研究者（女性）の協力を得て実験映像の撮影を行った．図 2 に撮影途中の光景およびカメラからの視点を示す．カメラには Panasonic HX-A500 を使用しているが，このカメラは十分に軽量であり，実験の邪魔にならないように配慮している．撮影における実験は普段実験を行うのと同じように行動してもらうことを依頼した．

撮影する実験について． 生化学分野の有名かつ基礎的な実験である PCR，ミニプレップ法，

図 2　(a) 実験映像の撮影風景．(b) 撮影途中の一人称カメラの視点．

DNA 抽出，アガロースゲル作成の 4 種類の実験に焦点を当てて映像を撮影した．1 つの実験あたり 8 の映像を撮影した．この際，DNA 抽出に関しては方法がエタノール沈殿法およびフェノールクロロホルム法の 2 通りの方法があるため，それぞれ 4 映像ずつ収録している．合計 32 の映像からなるデータセットである[3]．

映像の前処理について．いくつかの実験映像については，研究者が待機しなければならない手順が存在する（例：遠心分離機で 5 分遠心する）．こうした待ち時間には，研究者はカメラを外し待機しており，手順の内容とは関係がない．そのため，本研究では映像の前処理としてこうした時間を手動で取り除いている．

3.1.2 映像と言語の対応関係のアノテーション

著者のうち 1 名がアノテータとして映像と言語の対応関係のアノテーションを行った．実験を実施した研究室で使われているプロトコルはイラストなどを用いて視覚的に表現されており，明確な文書化はされていなかった．そのため，実験者に実験内容を口頭で説明してもらい，それを文字で書き起こしたものをプロトコルとして利用した．次に，プロトコルを動詞ごとに分割し，手順列を得た．例えば，“Invert 4 times to mix and add 10 μl of Alkaline Protease Solution.” という記述については，“Invert 4 times to mix” と “add 10 μl of Alkaline Protease Solution.” という文の 2 つの手順に分割した．この時，接続詞（ここでは and）については消去するよう指示した．ここでは，代名詞や省略については特別な処理を行ってはいない．次に，アノテータは映像を視聴し，各手順の開始時間，終了時間を決定してイベントのアノテーションを付与した．この時，開始時刻と終了時刻にできるだけ曖昧さが残らないように，手順対象の物体を持った時から，手順対象の物体を手放すまでの間をイベントとしてアノテーションした．表 1 に PCR のアノテーション結果の例を示す．節 1 で述べた通り，プロトコルの手順には，操作対象の物体名と操作に関する動詞を持ち，必要に応じて副詞や量，時間，数などの情報が追加で記述される．

3.1.3 プロトコル内に現れる物体の映像矩形アノテーション

さらに，別のアノテータがプロトコル中に現れる物体の映像矩形アノテーションを行った．このアノテータも著者のうちの 1 名である．映像から 4 秒ごとにフレームを抽出し[4]，各フレームに対して (1) 手が物体と触れている状態にあり，かつ (2) 触れている物体がプロトコル中に存

[3] 行う実験の内容は同じで，実験者も同じだが，初期状態は実験ごとに異なり，それに応じて行う動作なども映像ごとに異なる．節 3.2 で述べる通り，いくつかの手順はそれに応じて飛ばされている．こうした実際の実験の動作の違いにも富んだデータセットの構築のために，複数の映像を撮影している．

[4] 映像の全てのフレームに矩形アノテーションを行うことは高コストである．アノテーションコストを下げるために，映像から疎にサンプリングしてアノテーションを行うことが一般的である．Zhou ら (Zhou et al. 2019a) は映像あたり 10 フレームをサンプリングして矩形アノテーションを行っている．

在する場合において，その物体を矩形で囲い，物体名と紐づけるアノテーションを施している（図 3）．なお，これらの物体名は WLP データセット中の物体名に基づく固有表現に該当するものをプロトコルから抽出し利用している．この時，節 3.1.2 のアノテーション結果を参照しながら矩形アノテーションを行っている．なお，イベントのみに矩形アノテーションを行っている

手順文	開始	終了
add sterile distilled water	30	45
add primer1	64	99
add primer2	106	130
add template	149	173
add primeSTAR®Max Premix	190	238
set in DNA engine	260	266

表 1　PCR を対象としたアノテーション結果の一例．表の値は秒数を示す．なお，商標登録マーク ® は表では記載しているが，データセットの実際のプロトコルには含まれていない．

図 3　矩形アノテーションツールのアノテーション画面．ツールはブラウザからアクセスできるように開発されている．(A) にはプロトコル中に現れる物体が列挙されており，(B) にて矩形アノテーションを行う．(C) にて前後のフレームについて移動できる．

のではなく，映像全体にアノテーションを行っている点は注意されたい．映像中のイベント以外の区間においても物体をアノテーションしておくことで，幅広い応用（例：映像からのプロトコル生成）や映像の分析に利用することを想定している．

3.2 統計情報

次に，構築した BioVL2 データセットについて統計情報を言語側，映像側それぞれの観点から述べる．この結果より，両側面において BioVL2 データセットは多様な実験を収録したデータセットであることを示す．

3.2.1 プロトコル

表 2 に言語側の統計情報を示す．手順あたりの単語数と比較すると，手順数は実験ごとに大きく異なることが分かる．これは，BioVL2 データセットの手順の多様性を表している．中でも最も手順数が多かった実験はミニプレップ法であり，最も少ないのはフェノールクロロホルム法であることが分かる．物体の種類数についてはこれと同様の傾向を示しているが，動詞の種類数については最も種類が少ないのは PCR であった．言語の多様性を検証するために，ある実験にユニークな動詞と物体の種類数を調査した（表 3）．この結果と表 2 の右 2 列を比較すると，物体の種類数は一致するのに対し，動詞の種類数は著しく下がっている．このことから，

	手順数	手順あたりの単語数	物体の種類数	動詞の種類数
DNA 抽出				
フェノールクロロホルム法	4.0 (±0.0)	6.0 (±1.9)	2.0 (±0.0)	4.0 (±0.0)
エタノール沈殿法	9.0 (±0.0)	4.9 (±2.9)	4.5 (±1.7)	4.3 (±0.5)
PCR	6.0 (±0.0)	3.0 (±1.0)	6.0 (±0.0)	2.0 (±0.0)
アガロースゲル作成	10.3 (±0.4)	4.7 (±2.4)	5.5 (±0.5)	7.0 (±0.0)
ミニプレップ法	28.2 (±0.4)	6.4 (±2.5)	7.9 (±1.4)	8.1 (±1.2)

表 2 言語側の統計情報．各値について，平均と標準偏差を示している．ミニプレップ法とアガロースゲル作成について，同じ実験を行ってはいるが，一部の手順が状況に応じて飛ばされている．そのため，標準偏差が 0 にならない点は注意が必要である．

	ユニークな物体の種類数	ユニークな動詞の種類数
DNA 抽出		
フェノールクロロホルム法	2.0 (±0.0)	0.0 (±0.0)
エタノール沈殿法	4.5 (±1.7)	0.3 (±0.5)
PCR	6.0 (±0.0)	0.0 (±0.0)
アガロースゲル作成	5.5 (±0.5)	4.0 (±0.0)
ミニプレップ法	7.9 (±1.4)	3.0 (±1.1)

表 3 他の実験には現れない，ある実験にユニークな物体の種類数および動詞の種類数．

BioVL2 データセットでは動詞については全実験を通して共通のものが現れる一方，物体については実験ごとに固有の表現を持っていることが分かる．よって，物体に対する言語の多様性はあると言える．

3.2.2　実験映像

表 4 に各実験ごとの映像の長さ（秒数）について，図 4 に手順に紐づいたイベントの長さ（秒数）についての統計情報を示す．この結果から，映像やイベントの長さの観点からも，BioVL2 データセットは多様性に富んでいることが分かる．映像全体の長さが最も長いのはエタノール沈殿法であり（平均 399 秒），最も短いものは PCR である（平均 254 秒）．また，イベントごと

	映像の長さ（秒）
DNA 抽出	
フェノールクロロホルム法	269.4 (±58.6)
エタノール沈殿法	399.4 (±19.2)
PCR	254.6 (±18.1)
アガロースゲル作成	312.6 (±64.5)
ミニプレップ法	382.1 (±69.9)

表 4　実験ごとの映像の長さの統計情報．表の値は平均と標準偏差を示す．

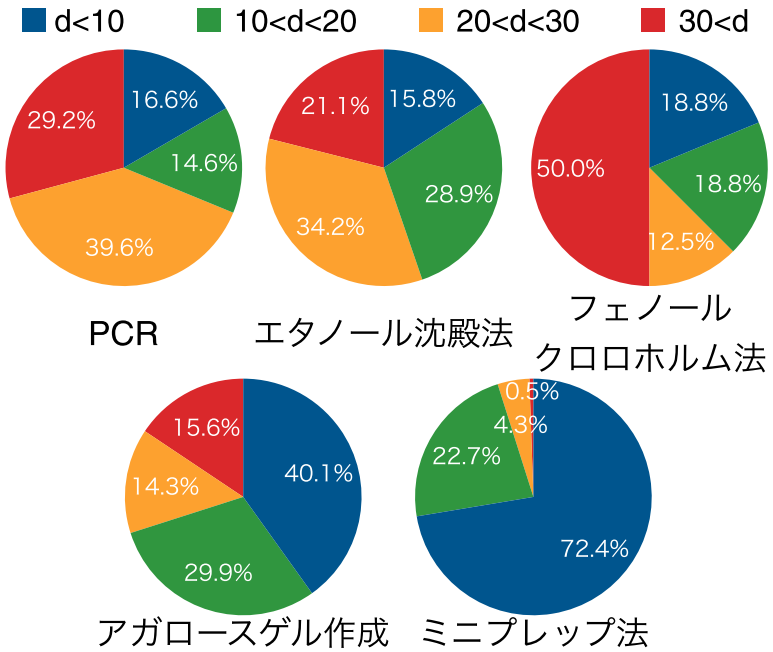

図 4　手順に紐づいたイベントの長さ（秒数）の割合．d は秒数を示す．

の長さに着目すると，ミニプレップ法はイベントの 72.4%は 10 秒以下の手順で構成されている一方，フェノールクロロホルム法は 50.0%のイベントが 30 秒以上の長いイベントを占めている．

3.2.3 矩形アノテーション

表 5 に少なくとも 1 つ以上矩形アノテーションを行ったフレームが何枚存在したかについて，表 7 にアノテーションした矩形の数についての統計情報を示す．また，表 6 に矩形アノテーションに紐づく物体を b-NE タグのカテゴリ種類に分類した上での統計情報を示す．b-NE タグの物

	アノテーションしたフレーム数	全フレーム数	割合
DNA 抽出			
フェノールクロロホルム法	195	263	74.1
エタノール沈殿法	265	394	67.3
PCR	322	498	64.7
アガロースゲル作成	282	614	45.9
ミニプレップ法	245	752	32.6

表 5 矩形アノテーションしたフレーム数の統計情報．

	Reagent の数	Location の数	Device の数
DNA 抽出			
フェノールクロロホルム法	1	1	0
エタノール沈殿法	6	0	0
PCR	5	0	1
アガロースゲル作成	2	3	1
ミニプレップ法	11	4	0
合計	25	8	2

表 6 矩形アノテーションに紐づく物体を b-NE タグのカテゴリ種類に分類した上での統計情報．実験ごとにユニークな物体を数え上げ報告している．b-NE タグのうち，物体に基づく固有表現には Reagent，Location，Device，そして Seal があるが，Seal は BioVL2 データセットの物体名に該当するものが存在しなかった．そのため，本アノテーションにおいては Reagent，Location，Device のみ数えた結果を示している．

	合計の矩形数	平均矩形数	最大矩形数
DNA 抽出			
フェノールクロロホルム法	68	1.0	1
エタノール沈殿法	130	1.0	2
PCR	176	1.0	1
アガロースゲル作成	372	1.1	2
ミニプレップ法	627	1.2	3

表 7 矩形アノテーションの統計情報．

体に基づく固有表現には Reagent，Location，Device，そして Seal があるが，Seal は BioVL2 データセットの物体名に該当するものが存在しなかった．そのため，Reagent，Location，Deivce のみ数えた結果を示している．この表より，Reagent が最も多く，ついで Location，そして Device が最も少ないことが分かる．表 5 より，アノテーションするフレーム数には実験ごとに違いがあることが分かる．最もアノテーションをしたフレームの割合が大きいのはフェノールクロロホルム法であり，最も小さいのはミニプレップ法である．表 7 に着目すると，概ねフレームあたり矩形を 1 つアノテーションしていることが分かる．また，ミニプレップ法はアノテーションするフレームの割合は最も低いものの，アノテーションしている矩形の数は最も多いことも分かる．

3.3　一致率の計算

アノテーション結果の品質を確かめるため，前節でアノテーションを実施した者とは別のアノテータにアノテーションを依頼し一致率を計算した．映像と言語の対応関係のアノテーションに関しては著者のうち 1 名が，映像矩形アノテーションについては著者とは別の 1 名が行った．この時の作業指示については全て節 3.1.2，節 3.1.3 と同様である．全ての映像に対してアノテーションを再度行うことはコストが高いため，実験ごとに 1 つランダムに映像を選択し，アノテーションを依頼した．以下，アノテーションの種類ごとに一致率を報告する．

3.3.1　映像と言語の対応関係のアノテーション

撮影した実験の経験者に映像と言語の対応関係のアノテーションを依頼し，その結果と節 3.1.2 の結果を比較して一致率を計算した．ここでは一致率の指標として 2 つのイベントの時間的な重なりを表現する temporal Intersection over Union (tIoU) を利用した．表 8 に一致率の結果を示す．この結果より，すべての実験において tIoU は平均 75%を超えていることが分かる．言語情報と映像を入力としたイベント検索課題 (Lei et al. 2020) において，tIoU が 0.7 を超えた時正解とみなして再現率を計算している．この基準から述べると，75%は高い値であると考えられ，アノテーションの品質は十分であることが分かる．

	tIoU
DNA 抽出	
フェノールクロロホルム法	88.9
エタノール沈殿法	91.7
PCR	99.4
アガロースゲル作成	82.2
ミニプレップ法	76.0

表 8　視覚と言語の対応関係アノテーションの一致率．表の値は tIoU の平均値を示す．

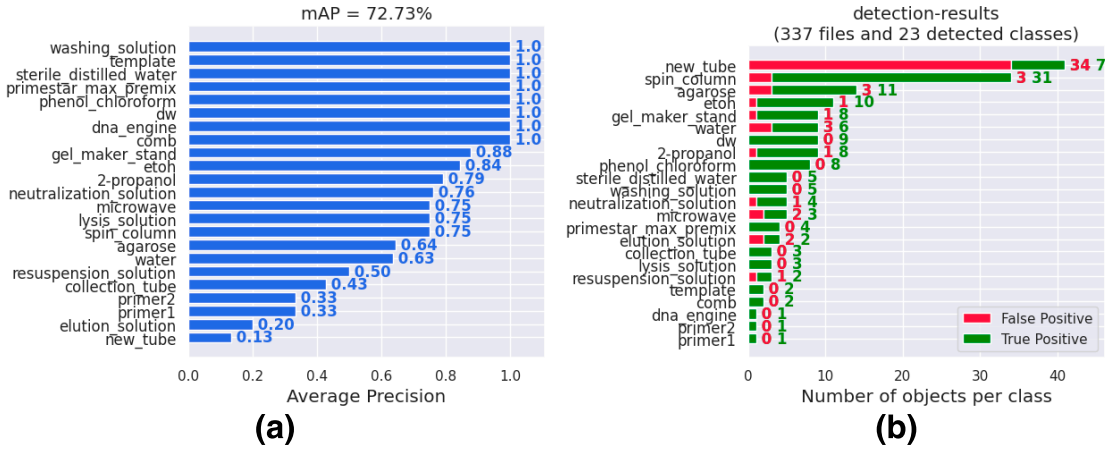

図 5　(a) 物体ごとの AP および mAP の結果．(b) 物体ごとの True positive（真陽性）と False positive（偽陽性）の数．

3.3.2　プロトコル内に現れる物体の映像矩形アノテーション

節 3.3.1 とは別のアノテータに映像への矩形アノテーションを依頼し，節 3.1.3 の結果と比較して一致率を計算した．ここでは，物体検出 (Ren et al. 2016) の評価尺度としてよく用いられる mean Average Precision (mAP) を用いた (Cartucho et al. 2018; Everingham et al. 2012)．新たにアノテーションした方を予測結果，節 3.1.3 の結果を正解とみなして mAP の計算を行った[5]．

図 5 に mAP の結果と真陽性と偽陰性の分布を示す．mAP は 72.73%となった．物体検出課題における最新モデル (Carion et al. 2020; Zhu et al. 2021) の mAP が 0.6 から 0.7 の間であり，実用的な性能に達していることを考えると，この値は高いと考えられ，アノテーションの品質は十分であることが分かる．一方で，多くの物体の AP が 0.7 を超えているものの AP が低い物体もあることが分かる．最も AP が低かったのは new_tube である．元々のアノテータはある手順以降の，新しい tube のみアノテーションを付与しているのに対し，新たにアノテーションを行った者はすべての tube にアノテーションを付与していることが原因であった．このように，複数のインスタンスを文脈に応じて呼び分ける必要があるが，new などの修飾語に応じた適切なアノテーション方法は今後検討の余地がある．

4　応用例：実験映像からのプロトコル自動生成

構築した BioVL2 データセットの応用課題の一例として，実験映像からプロトコルを自動生

[5] mAP の計算には，`https://github.com/Cartucho/mAP` を用いている（2022 年 5 月 11 日アクセス）．

成する課題に取り組む．BioVL2 データセットの映像の数は 32 と少なく，深層学習をもとにした映像キャプショニングモデル (Krishna et al. 2017; Zhou et al. 2018b) を適用することは困難である．少量の作業映像を対象とした手順書生成手法として，我々は Ushiku ら (Ushiku et al. 2017) の手法を活用する．この研究では料理映像を対象に，以下の 5 つのプロセスで手順書を生成する手法を提案している．(1) Faster-RCNN (Ren et al. 2016) によって各フレームに対して物体認識を行い，(2) 認識した物体の中からレシピ固有表現 (笹田 他 2015)（recipe-Named Entity，以下 r-NE）に該当する物体のみを抽出する．そして，(3) 連続する部分フレームの認識結果を統合して r-NE 列を得る．(4) あらかじめ大量のレシピで事前学習させておいた r-NE 集合から文を生成するモデル（(Ushiku et al. 2017) では LSTM を使用）を用いて，(3) の r-NE 列ごとに文を生成する．最後に，(5) 生成した文集合の中からレシピとして最も尤もらしい組み合わせをビタビアルゴリズムを用いて探索しレシピとして出力する．本手法の特徴的な点として，物体名以外の視覚情報を活用しない点にある．r-NE と対応する文の組み合わせは，固有表現認識器を用いれば Web 上から大量に収集し，文生成モデルの事前学習を行うことができる．こうして学習したモデルと物体認識器を併せて利用することで，少数の映像からでもレシピを生成することができる．

BioVL2 データセットでこの手法を適用するにあたり，以下の 2 点の変更を加えた．

- 本研究では (1) と (2) の処理を行う代わりに，矩形アノテーション結果を後段の処理に利用している．料理分野では矩形アノテーションがあれば映像中の物体認識はある程度行うことができるが (Hashimoto et al. 2014)，本研究で対象とする生化学分野においては物体を正しく認識することは困難であることが推察できる．例えば図 6 において Primer1 と Primer2 の認識を 1 枚のフレームから行うことは困難である．なぜなら，これらの物体は小さく認識が難しいだけでなく，外観も類似しているからである．映像を通して見

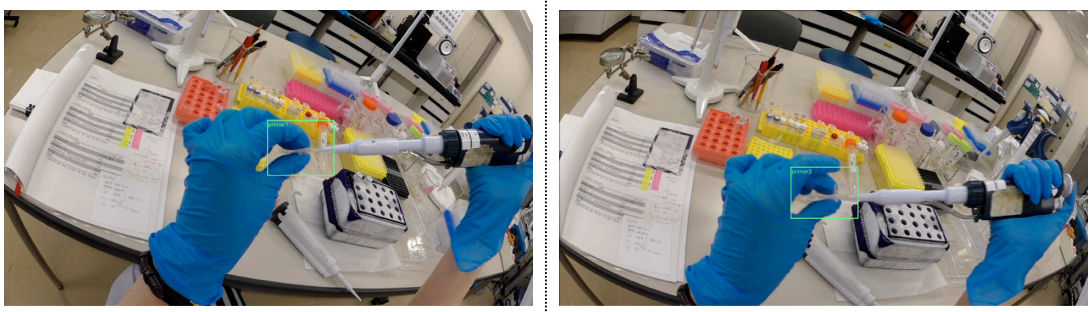

図 6　(1) Primer1 を操作している画像と (2) Primer2 を操作している画像.

ると作業者はこれらの物体を使い分けていることが分かるが，1 枚のフレームだけからこの識別を行うことは人にとってもほとんど不可能である．そのため，本研究では矩形アノテーションの結果を入力としてプロトコルを生成する課題に取り組む．この課題設定は，今後 QR コードなどを物体に貼り，何を持っているのかが機械的に推定できる状況を想定している．

- 文生成の事前学習モデルには LSTM ではなく Transformer (Vaswani et al. 2017) を用いる．Transformer は機械翻訳，文書要約 (Liu and Lapata 2019)，画像キャプショニング (Cornia et al. 2020) など多くの文生成課題において LSTM よりも良い性能を発揮しており，本課題においても有用であると考えられる．さらに，Transformer にコピー機構 (See et al. 2017) を導入するモデルも検討する．これは，入力の矩形アノテーションの物体情報を含む文を生成しやすくするためである．また，モデルの事前学習を行うデータセットとして，WLP データセット (Kulkarni et al. 2018) を活用する．

この変更を加えた手法の概要を図 7 に示す．以下，手法の詳細について順に説明する．ここで，矩形アノテーションが 1 つ以上存在するフレーム列を $\boldsymbol{F} = (f_1, f_2, \ldots, f_n, \ldots, f_{|\boldsymbol{F}|})$ とする．

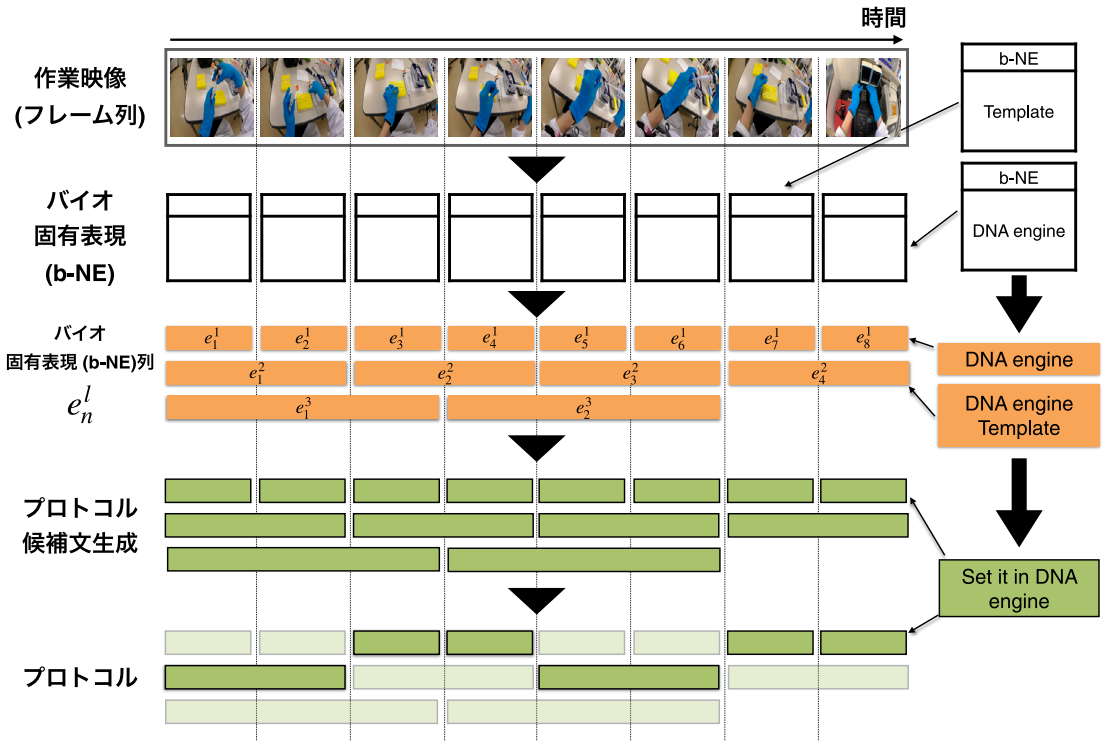

図 7　実験映像からのプロトコル生成手法．

我々の課題では，各フレームごとにアノテーション済みの物体名を用いて，b-NE 列を獲得する．例えば，図 7 では，DNA engine や (DNA engine, Template) が該当する．そして，事前学習した文生成モデルを用いて，b-NE 列ごとにプロトコル候補文を生成する（図 7 では，“Set it in DNA engine” が該当）．最後に，プロトコル候補文の中から，プロトコルとして最も尤もらしい組み合わせを Viterbi アルゴリズムを用いて探索し出力する．

4.1 b-NE 列の獲得

処理 (3) にならい，フレーム列の部分列 $f_n^{n+(l-1)}$（l は部分列の長さ）に対し，固有表現列を与える．与えられた固有表現列をバイオ固有表現 (b-NE) 列と呼ぶ．矩形アノテーションが 1 つ 1 つが b-NE であるとみなすと，n 番目のフレーム f_n は b-NE 集合 $\mathcal{E}_n$ を持つ．この時，部分フレーム列 $f_n^{n+(l-1)}$ に含まれる b-NE 集合は $\mathbf{e}_n^l \in \mathcal{E}_n \times \mathcal{E}_{n+1} \times \ldots \times \mathcal{E}_{n+(l-1)}$ として定式化する（$\times$ は集合のデカルト積）[6]．例えば，図 7 の $l = 1$ の例では，$\mathbf{e}_n^l$ は DNA engine のみ，$l = 2$ の例では，(DNA engine, Template) となる．部分列 l について，Ushiku らと同様に $l = 1, 2, 3$ の b-NE 列を構築する．

4.2 b-NE 列からのプロトコル候補文生成

4.2.1 文生成モデルの事前学習

次に，b-NE 列 $\mathbf{e}$ からプロトコルの候補文を生成する．これは，入力の b-NE 列に対して，その情報を保持したまま，適切な動詞や前置詞などを補完しつつ文を出力する課題である．図 8 に WLP データセットを用いた文生成モデルの事前学習についての概要を示す．モデルは Transformer をベースにした自己回帰型エンコーダデコーダモデルにコピー機構を加えている．入力として b-NE 列を `[CLS]` という特殊トークンを先頭に，`[SEP]` トークンで区切って結合した単語列 $\boldsymbol{X} = (x_1, x_2, \ldots, x_i, \ldots, x_{|\boldsymbol{X}|})$（$x_i$ は i 番目の単語）に対し，出力文を $\boldsymbol{Y} = (y_1, y_2, \ldots, y_j, \ldots, y_{|\boldsymbol{Y}|})$（$y_j$ は j 番目の単語）とする．Transformer エンコーダを $E(\cdot)$，Transformer デコーダを $D(\cdot)$ とすると，エンコーダの出力ベクトル列 $\boldsymbol{H} = (h_1, h_2, \ldots, h_i, \ldots, h_{|\boldsymbol{X}|})$，デコーダの j 番目の単語に対応する出力ベクトル o_j は以下のように計算する．

$$\boldsymbol{H} = E(\boldsymbol{X}) \tag{1}$$

$$o_j = D(\boldsymbol{Y}_{<j}, \boldsymbol{H}) \tag{2}$$

ここで，$\boldsymbol{Y}_{<j}$ は j 番未満の部分単語列を表す．図 8 にあるように，エンコーダの出力はデコーダの Source-target attention 層へ入力される．また，エンコーダとデコーダの単語の分散表現

[6] Ushiku らの手法では，Faster RCNN の物体検出結果をもとに，デカルト積を用いて r-NE 列を構築することで，検出誤りにも頑健にフレーム部分列を表現することを試みている．本研究では，人手でアノテーションした b-NE を後段の処理に使っているが，従来手法の b-NE 列の獲得方法を踏襲している．

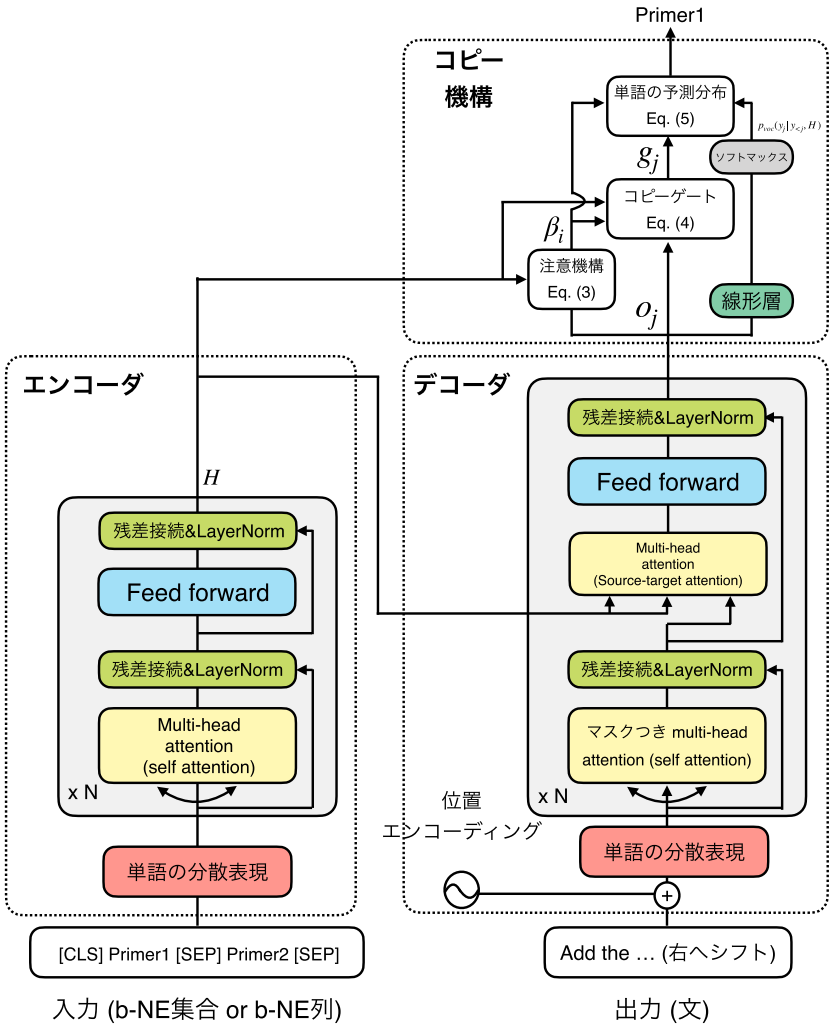

図 8　Transformer 文生成モデル．物体名を入力として対応する文を出力するように学習する．

層は別々に最適化しており，デコーダ側には位置エンコーディングが加算されているが，エンコーダ側は集合のエンコードであることを考慮してこれを加えていない点は注意されたい．

本研究の設定では，入力の物体名は出力に必ず含まれるため，それを陽にモデル化することでより正しく文生成を行えるようになると考えられる．これを加味するために，コピー機構をモデルへ導入する．エンコーダの出力 $\boldsymbol{H}$ とデコーダの出力 o_j を用いて，注意確率 β_j^i を以下のように計算する．

$$\beta_j^i = \frac{\exp\{(o_j)^\mathsf{T} W_c h_i\}}{\sum_k \exp\{(o_j)^\mathsf{T} W_c h_k\}}, \tag{3}$$

ただし，W_c は線形層を表す．次に，b-NE 列から単語を選択するか，語彙の単語を選択するかの確率をコピーゲート $g_j(0 \leq g_j \leq 1)$ を以下のように計算する．

$$g_j = \sigma(W_g[o_j; \sum_m \beta_j^m h_m] + b_g), \tag{4}$$

ここで，$[\cdot]$，$\sigma(\cdot)$，W_g，b_g はそれぞれベクトルの結合関数，シグモイド関数，線形層の重みおよびバイアスを表す．g_j を用いて，単語の予測分布 $p(y_j|y_{<j}, \boldsymbol{H})$ は以下のように計算する．

$$p(y_j|y_{<j}, \boldsymbol{H}) = (1 - g_j)p_{voc}(y_j|y_{<j}, \boldsymbol{H}) + g_j \sum_{i:x_i=y_j} \beta_j^i, \tag{5}$$

ここで，$p_{voc}(y_j|y_{<j}, \boldsymbol{H})$ は語彙中の単語 y_j の確率分布を表す．入力と出力の組 $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{D}$（$\mathcal{D}$ は学習データ全体を表す）について，以下の負の対数尤度 $\mathcal{L}(\theta)$ を最小化するように学習を行う．

$$\mathcal{L}(\theta) = -\sum_{(\boldsymbol{X},\boldsymbol{Y})\in\mathcal{D}} \log p(\boldsymbol{Y}|\boldsymbol{X}; \theta), \tag{6}$$

ここで，θ はモデルの学習可能なパラメータ全体を表す．

4.2.2 b-NE 列からのプロトコル候補文生成およびスコアの算出

WLP データセットで事前学習したモデルを用いて，節 4.1 で獲得した b-NE 列から文を生成する．また，後段のプロトコル出力の探索のために，生成したプロトコル候補文のそれぞれについて，生成した文の尤もらしさをスコアとして得る．具体的には，n 番目のフレーム，長さ l の b-NE 列 $\mathbf{e}_n^l$ に対し，以下の式に則ってスコアを求める．

$$Score(\mathbf{e}_n^l) = \prod_{i=1}^{N} p(d_i|d_1, d_2, \ldots, d_{k-1}; \mathbf{e}_n^l), \tag{7}$$

ここで，d_i は出力文の i 番目の単語を示し，N は単語列の長さを示す．

4.3 プロトコルの出力

生成した文集合の中から，プロトコルとして最も尤もらしい系列を選択しプロトコルを出力する．各文に対して与えるスコア $Score(\mathbf{e}_n^l)$ の和が最大となるようにフレーム列を選択する．ここで，2 つのヒューリスティックを導入する．第一に，1 人の研究者が 2 つの作業を同時に行うことはほとんど不可能であるため，プロトコル手順の候補に対応する部分フレーム列は重なってはならないこととする．第二に，同じ手順が複数回登場することはないとし，プロトコル中に一度登場した手順のスコアを 0 とする．この第二の仮定のもとでは，手順候補のスコアが変化しうるため，全探索を行ってプロトコルを出力するべきである．しかし，スコアの変化は同一文の生成時のみに限定され，頻繁に発生するものではないと考えられるため，ビタビアルゴ

リズム (Viterbi 1967) を用いて時間方向に沿って探索を行う．手順候補のスコアを高めるための経路を選択した結果を結合してプロトコルとして出力する．スコアが高ければ高いほど，プロトコルとして尤もらしいと考える．

5 評価

前節で説明した手法を BioVL2 データセットに適用してプロトコルを生成し評価を行った．以下では，実験設定について，利用したデータセット，モデルのハイパーパラメータなどの詳細設定，そして評価尺度の点から述べる．次に，生成したプロトコルの定量的，定性的な評価結果を報告する．評価の結果，我々はある程度正しくプロトコルを自動生成できていることを確認する．

5.1 実験設定

5.1.1 データセット

文生成モデルの学習において，b-NE 集合と文の対からなるデータセットが必要である．データセットのサイズの観点から BioVL2 データセットを学習に用いることは困難であるため，十分なサンプル数がある WLP データセット (Kulkarni et al. 2018) を学習に利用する．WLP データセットの各文には b-NE のアノテーションが行われているが，矩形アノテーションを付与した物体が，Reagent，Location，および Device のいずれかである点を踏まえ（表 6），本研究では特に Reagent，Location，Device の 3 つのタグを用いる．また，BioVL2 データセットの手順あたりの単語数は平均 10 単語以下である点を考慮し（表 2），(1) 文に前述した 3 つのタグのいずれかが含まれない，あるいは (2) 単語数が 20 を超える文については学習データと検証データから削除する．さらに，表 6 では Device は数が少ないことを加味して（Location の $\frac{1}{4}$），フィルタリングの (1) について，Reagent と Location のいずれかが含まれないケースも検討する．なお，フィルタリングの効果については，節 5.2 にて考察する．表 9 にて上記のフィルタリングを行わなかった場合と，行った場合の WLP データセットの統計情報を示す．このデータセットで学習した文生成モデルを用いて，BioVL2 データセットを評価データとして用いた．

5.1.2 詳細設定

文生成モデルの学習において，スペース区切りで文を単語に分割し，訓練データ内で頻度が 5 回以下のものは未知語として処理した．結果，語彙サイズは $1,827$ となった．Transformer の隠れ層のサイズは 768，層数は 2，Multi-head attention のヘッド数は 12 に設定した．位置エンコーディングには (Vaswani et al. 2017) と同様の方法を用いる．また，Transformer の学習には BERT (Devlin et al. 2019) の学習で行われているように，Adam (Kingma and Ba 2015) を学習

	全 b-NE タグを利用			Reagent, Location, Device のみ利用			Reagent と Location のみ利用		
フィルタリングなし	文数	単語数/文	b-NE 数/文	文数	単語数/文	b-NE 数/文	文数	単語数/文	b-NE 数/文
訓練データ	8,005	15.6	5.8	6,992	16.6	2.3	6,729	16.8	2.2
検証データ	2,709	15.5	5.8	2,353	16.6	2.4	2,261	16.7	2.2
フィルタリングあり									
訓練データ	5,984	11.5	4.6	4,996	12.1	1.9	4,761	12.2	1.8
検証データ	2,025	11.3	4.6	1,678	12.0	1.9	1,594	12.0	1.8

表 9 WLP データセットの統計情報．全ての b-NE タグを利用した場合，物体に基づく固有表現のみを利用した場合，そして Reagent と Location のみを利用した場合について，20 単語以上の文をフィルタリングするか，しないかの両ケースにおける統計を示している．データセットの分割は元の WLP データセットの分割に従っており，評価データは本研究においては利用していない．BioVL2 データセットを評価データとして用いているためである．

率を $\alpha = 0.0001$，$\beta_1 = 0.9$，$\beta_2 = 0.999$，L2 重み減衰率を 0.01 に設定し，最初の 5 エポックをウォームアップとした．最大エポックは 50 に設定し，バッチサイズは 16，WLP データセットの検証用データを用いて損失関数の値が最小になったモデルを評価に用いた．

5.1.3 評価尺度

文の自動評価尺度として広く利用される BLEU (Papineni et al. 2002)，METEOR (Banerjee and Lavie 2005)，ROUGE-L (Lin and Och 2004) を用いて評価を行った．BLEU の N については $N = 1, 2, 3, 4$ の値を評価結果として報告する[7]．なお，本研究で利用する手法によって生成されたプロトコルは必ずしも正解のプロトコルと同じ手順数になるとは限らないため，手順レベルでの評価は不可能である．そのため，本研究ではすべての手順を結合し，正解のプロトコルとの文書レベルでの評価結果を報告する．

5.2 定量的評価

文生成モデルとして，ベースラインモデル（入力の b-NE 列をプロトコルとして出力するモデル），Transformer のみのモデル，Transformer にコピー機構を加えたモデル（以下，Transformer+コピー機構と呼ぶ）との間で比較を行う．表 10 に文の自動評価尺度による結果を示す．この結果より以下の 4 点のことが明らかとなった．

第一に，正しいプロトコルの生成には文生成モデルの学習が必須であることである．ベースラインモデルに比べて Transformer モデルの性能が高いことから，このことが確認できる．第二に，WLP データセットは BioVL2 データセットの文生成に適用可能だということである．WLP データセットは生物学のプロトコルを収集したものであり，生化学分野のプロトコルを対象とした BioVL2 データセットとは分野がやや異なる（生化学は生物学の一分野）．しかし，BioVL2

[7] 評価を行うコードとして，`https://github.com/tylin/coco-caption` を用いた（2022 年 5 月 11 日アクセス）．

モデル	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L
ベースライン	3.0	2.0	1.1	0.5	6.2	6.1
全ての b-NE タグを学習に利用，単語数によるフィルタリングなし						
Transformer	15.0	8.7	4.2	2.0	11.1	22.3
Transformer+コピー機構	13.0	8.5	5.2	3.3	12.4	28.1
全ての b-NE タグを学習に利用，単語数によるフィルタリングあり						
Transformer	15.2	9.7	6.0	4.0	10.8	19.9
Transformer+コピー機構	13.3	7.9	3.7	1.8	12.2	27.6
Reagent, Location, Device のみ学習に利用，単語数によるフィルタリングなし						
Transformer	43.8	29.3	19.5	13.4	20.0	29.4
Transformer+コピー機構	36.8	26.9	20.0	14.8	18.5	**36.7**
Reagent, Location, Device のみ学習に利用，単語数によるフィルタリングあり						
Transformer	39.9	26.9	18.1	12.6	18.5	27.5
Transformer+コピー機構	42.5	30.6	21.3	15.5	20.7	32.7
Reagent と Location のみ学習に利用，単語数によるフィルタリングなし						
Transformer	36.8	25.8	18.2	13.3	19.5	24.9
Transformer+コピー機構	44.0	30.9	**21.8**	**16.4**	**21.7**	33.1
Reagent と Location のみ学習に利用，単語数によるフィルタリングあり						
Transformer	38.1	27.2	20.0	15.5	18.8	26.2
Transformer+コピー機構	**44.7**	**31.7**	**21.8**	15.2	21.1	32.2

表 10 文の自動評価尺度による評価結果．太字は最も結果が良い値を示す．

データセットの語彙の 83%は WLP データセットに現れ，両者のデータセットに大きな差はないことが分かる．第三に，Transformer+コピー機構は Transformer のみの文生成モデルよりも概ね良い性能を発揮することが分かる．これは，矩形アノテーションを付与した物体を正しく生成文に反映できているからと考えられる．第四に，学習に用いた b-NE タグの種類はモデルの性能に大きく影響を与えることである．全ての b-NE タグを学習に利用したモデル（フィルタリングなし）と，その他のモデルを比較すると後者のモデルは前者のモデルの性能を大きく上回った．また，Reagent，Location，Device のみを学習に用いたモデル（フィルタリングなし）と Reagent と Location のみを学習に用いたモデル（フィルタリングなし）を比較すると後者のモデルが概ね良い結果となった．これは，BioVL2 データセットに含まれる物体には Reagent と Location が大きい割合を占めるためと考えられる．一方，単語数によるフィルタリングには大きな性能の変化は見られなかった．

5.3 定性的評価

図 9 に Transformer のみ，Transformer+コピー機構，そして正解のプロトコルの比較結果を示す．節 1 で述べた通り，プロトコルとして正しい点であるためには，生成した各手順は正し

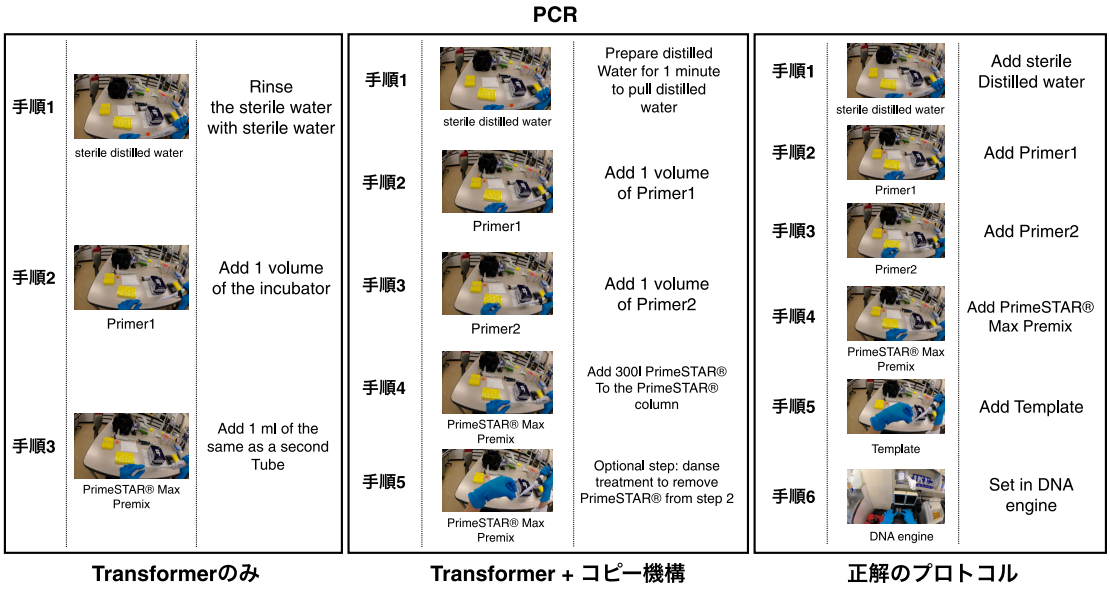

図 9　生成したプロトコルと正解のプロトコルの比較（実験は PCR）．利用したモデルは，表 10 中の「Reagent，Location，Device のみ学習に利用，単語数によるフィルタリングあり」．各プロトコルの中央列は選択したフレームを示す．ここで，各フレームの下には入力となる b-NE 列を表示している．商標登録マーク ® は論文中では図中には記載しているが，実際の生成文および正解の文には含まれていない．

く操作対象の物体を記述できており，かつ対応する動詞が正しくあらねばならない．また，必要に応じて副詞，量や時間に関する表現を追加で記述して作業者が再現できる形であらねばならない．Transformer のみのモデルの場合，正しく物体を予測できている手順も存在する一方で（例：手順 1 の sterile water），概ね物体の言語化において失敗していることが分かる．例えば，Primer1，Primer2，PrimeSTAR® などの単語については正しく言及できていない．一方で，コピー機構を Transformer に加えることにより，これらの物体の言及にある程度正しく生成できていることが分かる（例：手順 2 における Primer1，手順 3 における Primer2，手順 4 における PrimeSTAR® など）．以上の結果を踏まえると，プロトコルとして必要な要件の中で，物体に関する生成という点では，コピー機構を加えることが有用であることが分かる．

限界点．コピー機構を入れることである程度正しくプロトコルを生成できるようになった一方で，このモデルには 2 点の限界点がある．第一に，動詞の生成である．図 9 の手順 1 において Prepare という動詞を生成しているが，正確には Add を生成するべきである．本研究の手法には視覚情報は物体以外入力に与えていないため，動詞を正しく生成できるかは文生成モデルに依存する．この点は信頼性の高いプロトコル生成の点で提案手法が不十分な点の 1 つであろう．

第二に，全ての物体が必ずしも生成したプロトコル中に含まれるとは限らない点である．図

9 の正解の手順 6 において “DNA engine” という単語が文中に存在するが，生成したプロトコルの中にはこれが含まれていない．与えた物体を全て言語化することは正しくプロトコルの生成する上で重要であり，この点は今後の課題である．

第三に，数や量についての正しい生成が挙げられる．図 9 中では 1 volume や 1 minutes，300l などの量や時間に関する表現が生成されていることが分かるが，これらの量は実際には適切ではない[8]．これも，映像の情報を活用していないことの明確な限界点であり，動詞の生成と同じく正しい時間や量が生成されるかどうかは生成モデルに依存する．この量に関する記述も正しいプロトコル生成の点で提案手法の不十分な点であろう．今後，映像の情報を活用した文生成を行っていく必要がある．

5.4 物体と動詞の自己相互情報量の観点からの文生成モデルの挙動の考察

図 9 中の Transformer+コピー機構において，手順 2 や手順 3 において，物体のみしか与えていないにも関わらず，正しく動詞 “Add” を生成できている．前述した通り，これは文生成モデルに依存しているが，ある物体が決まれば正しい動詞を生成できるといったことは起こり得るだろうか．この点を明らかにするために，WLP データセットの訓練データにおける頻度が最も高い 10 の動詞，および物体とその出現頻度，および自己相互情報量 (PMI) が高い物体，および動詞の上位 5 件を計算した．

$$\mathrm{PMI}(\text{物体，動詞}) = \log_2 \frac{P(\text{物体，動詞})}{P(\text{物体})P(\text{動詞})} = \log_2 \frac{C(\text{物体，動詞})N}{C(\text{物体})C(\text{動詞})} \tag{8}$$

ここで，$C(\cdot)$ はある物体か動詞の単語の数，またはその両方が共起する文の数を示す．N は訓練データ中の全ての単語数を示す．表 11，12 にその結果を示す．その結果，ある物体に対して

動詞	動詞の頻度	PMI が高い物体
add	949	dna stripping solution (6.74), precipitation solution (6.74), chloroform (6.65), stand (6.60), genomic lysis buffer (6.49)
incubate	496	dark (6.84), assays (6.80), sections (6.67), nanobeads (6.35), primary antibody (6.29)
mix	374	dna stripping solution (8.08), precipitation solution (7.94), components (7.93), choloform (7.93), contents (6.99)
remove	346	upper phase (8.05), forceps (7.45), petri dish (7.32), xylenes (7.19), upper reservoir (6.81)
centrifuge	339	centrifuge tube (8.22), centrifuge (8.20), upper phase (8.08), cell strainer (7.57), cell pellet (7.36)
place	293	magnetic rack (7.72), petri dish (7.56), top (7.40), magnetic stand (7.30), magnet (7.22)
wash	283	ethanol wash (8.45), wash buffer (8.45), intracellular staining perm wash buffer (8.45), dna pellet (8.21), each well (7.94)
transfer	223	pcr tube (7.74), microfuge tube (7.69), isopropanol (7.55), cuvette (7.50), aqueous dna (7.09)
discard	220	flow-through (8.40), cell pellet (8.07), supernatants (7.67), spin cartridge (7.57), supernatant (7.52)
resuspend	187	cell strainer (8.43), mojosort (8.28), intracellular staining perm wash buffer (8.20), te (7.81), cell staining buffer (7.78)

表 11 WLP データセットの訓練データにおける，頻度が最も高い 10 の動詞とその出現頻度および PMI が高い物体上位 5 件．低頻度な単語は PMI が過剰に高くなる傾向があるため，頻度が 11 以上の物体の上位 5 件を報告している．

[8] PCR 実験においてプロトコルには量は記述されていない．PCR はキットによって試薬の量などがあらかじめ決められているため，実験者はそれを順に操作していくだけとなっている．

PMI の値が極端に高い物体が存在するわけではなく，上位 5 件は概ね同程度の値を示した．動詞からみた物体についても同様である．このことから，ある動詞あるいは物体に対し，特定の物体や動詞が一意に定まることはないと考えられる．ではなぜ図 9 にて正しい動詞を生成できたのであろうか．1 つの考えられる仮説として，モデルは高頻度な動詞を生成しやすいことがある．実験において，Primer1 や Primer2 は文生成時に未知語として処理されている[9]．しかし，add はもっとも頻度が高い動詞であるので，図 9 ではこれを出力したものではないかと考えられる．

5.5 プロトコルの手順数と性能についての考察

表 13 に生成したプロトコルと正解のプロトコルの間の実験ごとの手順数と文の自動評価尺度

物体	物体の頻度	PMI が高い動詞
tube	435	flicking (7.71), placing (7.48), disturbing (7.35), inverting (7.31), insert (7.12)
cells	406	scale up (7.82), seperating (7.72), lyse (7.67), working (7.58), flicking (7.44)
sample	309	load (6.88), insert (6.62), process (6.62), determine (6.47), cool (6.47)
supernatant	298	pour off (7.70), dissociate (7.56), discard (7.43), save (6.95), decent (6.85)
ethanol	171	disturbing (8.69), pipette off (8.53), invert (7.42), immerse (7.30), removed (7.21)
plate	170	plate (9.22), dissociate (8.95), spread (8.37), read (8.32), seal (8.07)
dna	162	precipitate (8.72), invert (8.20), elute (7.91), running (7.55), store (7.50)
ice	162	chill (8.55), thaw (8.55), kept (8.28), keep (8.00), incubating (7.87)
tubes	157	disturbing (8.59), precipitate (7.71), aliquot (7.67), label (7.48), clamp (6.87)
beads	130	disturbing (9.28), vortexing (7.92), seperate (7.60), spin (7.38), pipet (7.32)

表 12 WLP データセットの訓練データにおける，頻度が最も高い 10 の物体とその出現頻度および PMI が高い動詞上位 5 件．低頻度な単語は PMI が過剰に高くなる傾向があるため，頻度が 11 以上の動詞の上位 5 件を報告している．

	生成したプロトコルの手順数	正解プロトコルの手順数	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L
DNA 抽出								
フェノールクロロホルム法	3.0 (±0.0)	4.0 (±0.0)	34.5	23.5	16.0	11.2	18.2	39.1
エタノール沈殿法	6.3 (±1.5)	9.0 (±0.0)	28.7	15.5	6.7	0.0	15.9	30.0
PCR	5.0 (±0.0)	6.0 (±0.0)	23.5	15.0	8.5	0.0	22.3	33.5
アガロースゲル作成	5.0 (±1.2)	10.3 (±0.4)	33.2	21.4	10.0	0.0	13.8	24.7
ミニプレップ法	10.9 (±2.7)	28.2 (±0.4)	49.4	36.6	26.8	19.8	24.5	36.1

表 13 生成したプロトコルと正解のプロトコルの実験ごとの手順数，および文の自動評価尺度の結果．なお，ここで用いたモデルは表 10 中の「Reagent，Location，Device のみ学習に利用，単語数によるフィルタリングあり」である．

[9] 入力が語彙中に存在しない場合，モデルには未知語として与えられる．しかし，コピー機構は入力をそのまま出力文に含められるため，未知語であっても対応する単語をコピーして出力することが可能である．複数未知語が存在する場合に特別な処理を施してはいない．なお，BioVL2 中の物体名のうち，未知語を含む物体の割合は 34.3%(= 12/35) であった．

の結果を示す．最も手順数の差が最も小さいのは PCR とフェノールクロロホルム法，最も差が大きかったのはミニプレップ法であった．手順数の差は小さい方が理想的である．一方で，文の自動評価尺度による評価では手順数の差が大きいミニプレップが最も良い性能を示している．生成したプロトコルの手順数が多くなるほど，文書で見た時の単語数が増加し，正解と比較して一致する n-gram が増加する．このことが，文書レベルでの評価を行った時に手順数が最も多いミニプレップが，正解との手順数に差はあれど評価結果が高くなる理由であると考えられる．今後，プロトコル生成の評価として，手順数の差も考慮に入れた評価，正しい動詞や物体が生成したプロトコル中に現れるかどうかといった，プロトコルの質をその要件と照らし合わせた評価を行う仕組みを検討する必要がある．

6 おわりに

本研究は生化学分野における一人称の実験映像データセットである BioVL2 データセットを構築した．BioVL2 データセットは生化学における 4 種類の基本的な実験に対し，それぞれ 8 動画撮影した合計 32 映像からなるデータセットであり，言語アノテーションとして (1) 視覚と言語の対応関係のアノテーション，(2) プロトコル中に現れる物体の矩形アノテーションの 2 種類のアノテーションを付与している．構築したデータセットを用いて，応用課題として実験映像からプロトコルを自動生成する課題に取り組んだ．その結果，ある程度正しくプロトコルを生成できることを確認した．今後，BioVL2 データセットは研究用途に限り公開予定である．このデータセットを通して生化学分野における人工知能技術の発展が進むことを強く望む．

謝 辞

本研究は JSPS 科研費 JP21J20250，JP20H04210，JP21H04910 の助成を受けたものです．また，本研究を進めるにあたって，オムロンサイニックエックス株式会社の牛久 祥孝氏からは有益なコメントを頂きました．本論文に内容の一部は The 4th Workshop on Closing the Loop Between Vision and Language (CLVL21) で発表したものです (Nishimura et al. 2021b).

付録

A WLP データセットで定義されたタグの種類

表 14 に WLP データセットで定義された全てのタグ，それらが属するカテゴリ，そしてそ

タグの種類	タグのカテゴリ	タグの説明
Action	動作を示す固有表現	動詞，動作
Reagent	物体に基づく固有表現	薬品名
Location		試薬やその他物体を入れるための容器（ガラス器具など）
Device		容器としてだけでなく，特定の機能を持つ機械（遠心分離機など）
Seal		物体に付与したシールや蓋など
Amount	測量に基づく固有表現	物体の絶対量
Concentration		混合物の 2 つ以上の量の相対的な比率
Time		ある手順の特定の動作の所要時間
Temperature		温度（華氏や摂氏の表現）
Method		ある動詞に関連して実施される手法を簡潔に表現した名詞
Speed		一般的には，遠心分離機の 1 分あたりの回転数
Numerical		時間や温度に当てはまらない，単位を伴わない数字
Generic-Measure		時間や温度に当てはまらないが，単位を伴う数字（例：1mL）
Size		物体の寸法を表す尺度（例：長さや面積）
Measure-Type		測定の種類を示す汎用タグ（例：volume や density）
pH		溶液の酸性度またはアルカリ度を示す尺度
Modifier	品詞に基づく固有表現	副詞
Mention		過去に登場した物体を指す単語

表 14　WLP データセットで定義されたタグ.

の説明を示す．元論文 (Kulkarni et al. 2018) の定義によると，タグのカテゴリは動作を表すもの（表中の Action），物体に基づく固有表現を示すもの（表中の Reagent から Seal まで），測量に基づく固有表現（表中の Amount から pH まで），そして品詞に基づく固有表現 (Modifier と Mention) に分類されている.

B　tIoU の計算

イベント A，イベント B 開始時間と終了時間の組からなり，その区間 A，B とする．2 つのイベントの tIoU は以下の式で計算できる.

$$\mathrm{tIoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{9}$$

ここで，$|A \cap B|$ と $|A \cup B|$ はそれぞれイベント区間の積集合と和集合を表す.

参考文献

Baker, M. (2016). “1,500 Scientists Lift The Lid on Reproducibility.” *Nature*, **533**, pp. 452–454.

Banerjee, S. and Lavie, A. (2005). “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation With Human Judgments.” In *Proceedings of ACL Workshop IEEMMTS*, pp. 65–72.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). “The Mathematics of Statistical Machine Translation: Parameter Estimation.” *Computational Linguistics*, **19**, pp. 263–311.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). “End-to-End Object Detection with Transformers.” In *Proceedings of ECCV*, pp. 213–229.

Cartucho, J., Ventura, R., and Veloso, M. (2018). “Robust Object Recognition Through Symbiotic Deep Learning In Mobile Robots.” In *Proceedings of IROS*, pp. 2336–2341.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). “Meshed-Memory Transformer for Image Captioning.” In *Proceedings of CVPR*, pp. 10578–10587.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., and Wray, M. (2018). “Scaling Egocentric Vision: The EPIC-KITCHENS Dataset.” In *Proceedings of ECCV*, pp. 753–771.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of NAACL-HLT*, pp. 4171–4186.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A. (2012). “The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.” `http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html`.

Hashimoto, A., Sasada, T., Yamakata, Y., Mori, S., and Minoh, M. (2014). “Kusk Dataset: Toward A Direct Understanding of Recipe Text and Human Cooking Activity.” In *Proceedings of Ubicomp*, pp. 583–588.

Kiddon, C., Ponnuraj, G. T., Zettlemoyer, L., and Choi, Y. (2015). “Mise en Place: Unsupervised Interpretation of Instructional Recipes.” In *Proceedings of EMNLP*, pp. 982–992.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). “GENIA Corpus - A Semantically Annotated Corpus for Bio-textmining.” *Bioinformatics*, **19**, pp. i180–i182.

Kingma, D. P. and Ba, J. (2015). “Adam: A Method for Stochastic Optimization.” In *Proceedings of ICLR*.

Knafou, J., Naderi, N., Copara, J., Teodoro, D., and Ruch, P. (2020). “BiTeM at WNUT 2020 Shared Task-1: Named Entity Recognition over Wet Lab Protocols using an Ensemble of Contextual Language Models.” In *Proceedings of WNUT*, pp. 305–313.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Niebles, J. C. (2017). “Dense-Captioning Events

in Videos." In *Proceedings of ICCV*, pp. 706–715.

Kulkarni, C., Xu, W., Ritter, A., and Machiraju, R. (2018). "An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols." In *Proceedings of NAACL-HLT*, pp. 97–106.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." In *Proceedings of AAAI*, pp. 282–289.

Lei, J., Yu, L., Berg, T. L., and Bansal, M. (2020). "TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval." In *Proceedings of ECCV*, pp. 447–463.

Lin, C.-Y. and Och, F. J. (2004). "Automatic Evaluation of Machine Translation Quality using Longest Common Subsequence and Skip-bigram Statistics." In *Proceedings of ACL*, pp. 605–612.

Liu, Y. and Lapata, M. (2019). "Hierarchical Transformers for Multi-Document Summarization." In *Proceedings of ACL*, pp. 5070–5081.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). "HowTo100M: Learning a Text-video Embedding by Watching Hundred Million Narrated Video Clips." In *Proceedings of ICCV*, pp. 2630–2640.

Naim, I., Song, Y., Liu, Q., Kautz, H., Luo, J., and Gildea, D. (2014). "Unsupervised Alignment of Natural Language Instructions with Video Segments." In *Proceedings of AAAI*, pp. 1558–1564.

Naim, I., Song, Y. C., Liu, Q., Huang, L., Kautz, H., Luo, J., and Gildea, D. (2015). "Discriminative Unsupervised Alignment of Natural Language Instructions with Corresponding Video Segments." In *Proceedings of NAACL*, pp. 164–174.

Nishimura, T., Hashimoto, A., Ushiku, Y., Kameko, H., and Mori, S. (2021a). "State-aware Video Procedural Captioning." In *Proceedings of ACMMM*, pp. 1766–1774.

Nishimura, T., Sakoda, K., Hashimoto, A., Ushiku, Y., Tanaka, N., Ono, F., Kameko, H., and Mori, S. (2021b). "Egocentric Biochemical Video-and-Language Dataset." In *Proceedings of CLVL*, pp. 3129–3133.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation." In *Proceedings of ACL*, pp. 311–318.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, pp. 1137–1149.

笹田鉄郎, 森信介, 山肩洋子, 前田浩邦, 河原達也 (2015). レシピ用語の定義とその自動認識のためのタグ付与コーパスの構築. 自然言語処理, **22**, pp. 107–131. [T. Sasada et al. Definition of Recipe Terms and Corpus Annotation for their Automatic Recognition. Journal of Natural

Language Processing, 22, pp. 107–131.].

See, A., Liu, P. J., and Manning, C. D. (2017). "Get To The Point: Summarization with Pointer-Generator Networks." In *Proceedings of ACL*, pp. 1073–1083.

Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., and Yao, A. (2022). "Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities." In *Proceedings of CVPR*, pp. 21096–21106.

Singh, J. and Wadhawan, A. (2020). "PublishInCovid19 at WNUT 2020 Shared Task-1: Entity Recognition in Wet Lab Protocols using Structured Learning Ensemble and Contextualised Embeddings." In *Proceedings of WNUT*, pp. 273–280.

Sohrab, M. G., Duong, K., Miwa, M., and Takamura, H. (2020). "mgsohrab at WNUT 2020 Shared Task-1: Neural Exhaustive Approach for Entity and Relation Recognition Over Wet Lab Protocols." In *Proceedings of WNUT*, pp. 290–298.

Ushiku, A., Hashimoto, H., Hashimoto, A., and Mori, S. (2017). "Procedural Text Generation from an Execution Video." In *Proceedings of IJCNLP*, pp. 326–335.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is All You Need." In *Proceedings of NeurIPS*, pp. 5998–6008.

Viterbi, A. (1967). "Error Bounds for Convolutional Codes and An Asymptotically Optimum Decoding Algorithm." *IEEE Transactions on Information Theory*, **13**, pp. 260–269.

Wang, S., Zhao, W., Kou, Z., Shi, J., and Xu, C. (2021). "How to Make a BLT Sandwich? Learning VQA Towards Understanding Web Instructional Videos." In *Proceedings of WACV*, pp. 1130–1139.

Wang, W., Wang, Y., Chen, S., and Jin, Q. (2019). "YouMakeup: A Large-Scale Domain-Specific Multimodal Dataset for Fine-Grained Semantic Comprehension." In *Proceedings of EMNLP-IJCNLP*, pp. 5133–5143.

Xu, J., Mei, T., Yao, T., and Rui, Y. (2016). "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language." In *Proceedings of CVPR*, pp. 5288–5296.

Zhou, L., Kalantidis, Y., Chen, X., Corso, J. J., and Rohrbach, M. (2019a). "Grounded Video Description." In *Proceedings of CVPR*, pp. 6578–6587.

Zhou, L., Louis, N., and Corso, J. J. (2019b). "Weakly-Supervised Video Object Grounding from Text by Loss Weighting and Object Interaction." In *Proceedings of BMVC*.

Zhou, L., Xu, C., and Corso, J. J. (2018a). "Towards Automatic Learning of Procedures From Web Instructional Videos." In *Proceedings of AAAI*, pp. 7590–7598.

Zhou, L., Zhou, Y., Corso, J. J., Socher, R., and Xiong, C. (2018b). "End-to-End Dense Video Captioning With Masked Transformer." In *Proceedings of CVPR*, pp. 8739–8748.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). "Deformable DETR: Deformable Transformers for End-to-End Object Detection." In *Proceedings of ICLR*.

略歴

西村　太一：2019 年九州大学芸術工学部卒業，2020 年京都大学大学院情報学研究科修士課程修了．現在同大学博士課程．修士（情報学）．マルチメディア，自然言語処理，コンピュータビジョンの研究に従事．言語処理学会学生会員．学術振興会特別研究員 (DC1).

迫田航次郎：2020 年神戸大学工学部卒業．2022 年京都大学大学院情報学研究科修士課程修了．修士（情報学）．言語処理学会学生会員．

牛久　　敦：2017 年京都大学大学院情報学研究科修士課程修了．修士（情報学）．

橋本　敦史：2005 年京都大学工学部卒業，2006 年経産省 Vulcanus in Europe プログラム国費奨学生．2013 年京大大学院情報学研究科にて博士（情報学）取得．現在オムロンサイニックエックス株式会社研究員．主に，料理や組立作業を対象として，未来予測に基づく人と機械のインタラクションに関する研究などに従事．IEEE, IEICE, IPSJ 各会員．

奥田奈津子：2012 年神戸大学発達科学部卒業．2016 年より大阪医科大学（現 大阪医科薬科大学）医学部生命科学講座生理学教室に研究補助員として入職．現在，主たる実験動物ゼブラフィッシュの維持管理と分子生物学研究補助に従事．

小野富三人：1991 年東京大学医学部卒業．国立国際医療研究センターの研修医を経て，1996 年東京大学医学部博士課程修了．学術振興会特別研究員，ニューヨーク州立大学ポスドク，フロリダ大学助教授，NIH 室長をへて 2014 年より大阪医科大学（現大阪医科薬科大学）生理学教室教授．神経系を中心とした生理学の研究に従事．日本生理学会，北米神経科学会会員．

亀甲　博貴：2018 年東京大学大学院工学研究科博士課程修了．博士（工学）．同年より京都大学学術情報メディアセンター助教．自然言語処理，ゲーム AI 等に関する研究に従事．言語処理学会，情報処理学会各会員．

森　　信介：1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了．同年日本アイ・ビー・エム株式会社入社．2007 年より京都大学学術情報メディアセンター准教授．2016 年同教授．現在に至る．計算言語学ならびに自然言語処理の研究に従事．博士（工学）．1997 年情報処理学会山下記念研究賞受賞．2010 年，2013 年情報処理学会論文賞受賞．2010 年第 58 回電気科学技術奨励賞．言語処理学会，情報処理学会，日本データベース学会各会員．

（2022 年 5 月 13 日　受付）
（2022 年 7 月 26 日　再受付）
（2022 年 8 月 30 日　採録）